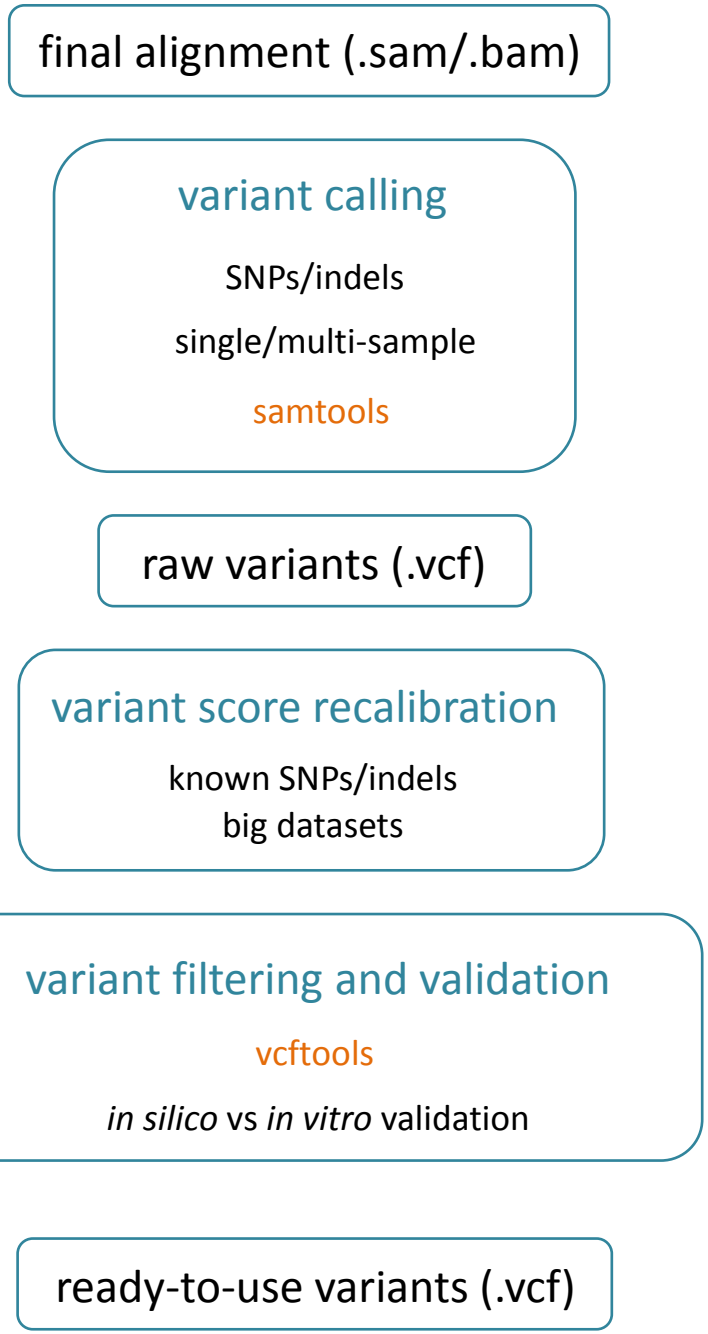
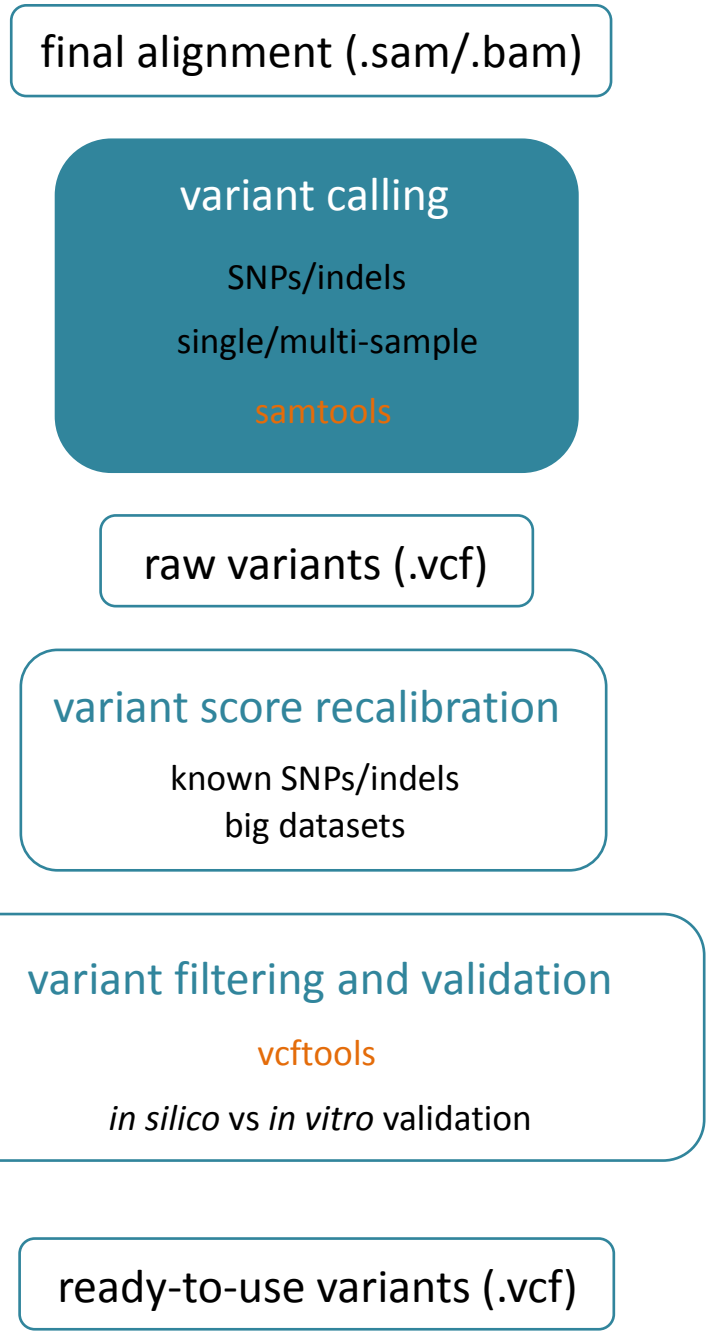

final alignment (.sam/.bam)
variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
variant filtering and validation
vcftools
in silico vs *in vitro* validation
ready-to-use variants (.vcf)


final alignment (.sam/.bam)
variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
variant filtering and validation
vcftools
in silico vs *in vitro* validation
ready-to-use variants (.vcf)

variant calling - SNPs

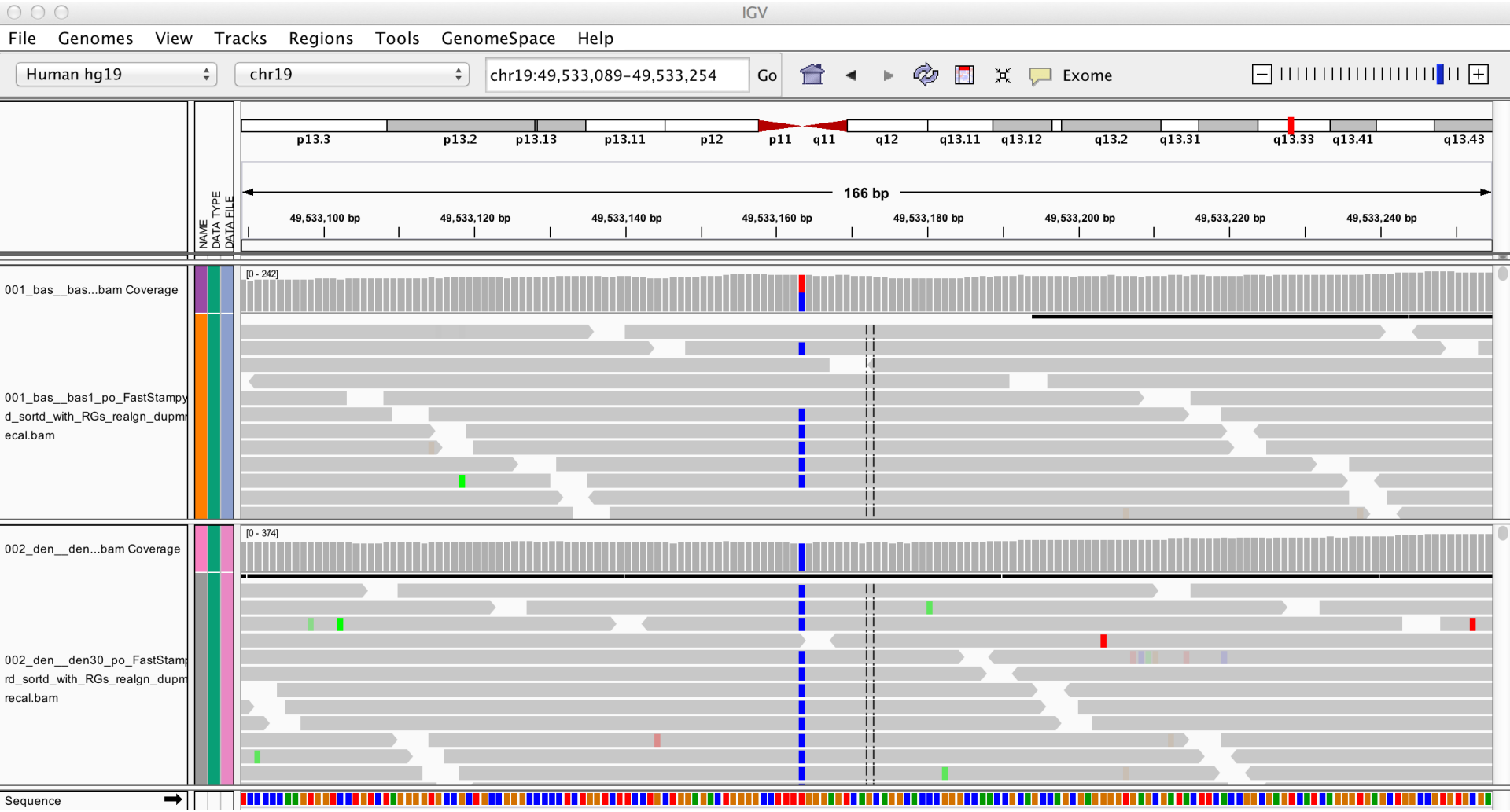
Examine the bases aligned to position and look for differences

SNP calling vs genotyping
Homozygous vs Heterozygous SNPs

Factors to consider:

- Base call qualities of each supporting base
- Mapping qualities of the reads supporting the SNP (increased read length or paired-end help MQ scores)
- Sequencing depth
- Individual vs multi-sample calling

variant calling – visualisation in IGV



variant calling – VCF file

Standardised format for storing DNA polymorphism data
- SNPs, indels, SV
- Rich annotations

Can be indexed for fast data retrieval of variants from a range of positions

Can store variant information over many samples

Record meta-data about the site
- dbSNP accession, filter status

Very flexible
- Tags can be introduced to describe new types of variants
- Different VCF files may contain different information/annotations

Two sections:
- Header
- Data

variant calling – VCF file

Header
lines starting with ##: arbitrary number of meta-information lines
line starting with #: column definition – mandatory columns include:

CHROM	chromosome
POS	position of the start of the variant
ID	unique identifier of the variant (e.g. rs number for SNPs)
REF	reference allele
ALT	comma separated list of alternate non-reference alleles
QUAL	phred-scaled quality score
FILTER	site filtering information
INFO	user extensible annotation (e.g. samtools and GATK may differ in this)

samples follow

Data
one line per site (all columns described above per line); useful information per site and per sample

variant calling – VCF file

```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than i
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FILTER=<ID=StrandBias,Description="Min P-value for strand bias (INFO/PV4) [0.0001]">
##FILTER=<ID=EndDistBias,Description="Min P-value for end distance bias (INFO/PV4) [0.0001]">
##FILTER=<ID=MaxDP,Description="Maximum read depth (INFO/DP or INFO/DP4) [10000000]">
##FILTER=<ID=BaseQualBias,Description="Min P-value for baseQ bias (INFO/PV4) [0]">
##FILTER=<ID=MinMQ,Description="Minimum RMS mapping quality for SNPs (INFO/MQ) [10]">
##FILTER=<ID=Qual,Description="Minimum value of the QUAL field [10]">
##FILTER=<ID=MinAB,Description="Minimum number of alternate bases (INFO/DP4) [2]">
##FILTER=<ID=VDB,Description="Minimum Variant Distance Bias (INFO/VDB) [0.015]">
##FILTER=<ID=GapWin,Description="Window size for filtering adjacent gaps [3]">
##FILTER=<ID=MapQualBias,Description="Min P-value for mapQ bias (INFO/PV4) [0]">
##FILTER=<ID=SnpGap,Description="SNP within INT bp around a gap to be filtered [10]">
##FILTER=<ID=MinDP,Description="Minimum read depth (INFO/DP or INFO/DP4) [2]">
##FILTER=<ID=RefN,Description="Reference base is N []">
##FILTER=<ID=HWE,Description="Minimum P-value for HWE (plus F<0) (INFO/HWE and INFO/G3) [0.0001]">
##source_20130519.1=/usr/bin/vcf-annotate -f +
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  60A-Sc-DBVPG6044
I       2709    .       G       T       74.1    PASS    DP=8;VDB=0.0321;AF1=1;AC1=2;DP4=0,0,6,1;MQ=60;FQ=-48    GT:PL:GQ        1/1:107,21,0:39
I       2825    .       G       C       73.3    PASS    DP=5;VDB=0.0302;AF1=1;AC1=2;DP4=0,0,1,4;MQ=60;FQ=-42    GT:PL:GQ        1/1:106,15,0:27
I       2875    .       TAA     TAAA    96.5    PASS    INDEL;DP=11;VDB=0.0321;AF1=0.5;AC1=1;DP4=3,1,1,5;MQ=60;FQ=44.5;PV4=0.19,1,1,0.041      GT:
I       2891    .       G       T       156     PASS    DP=12;VDB=0.0280;AF1=1;AC1=2;DP4=0,0,5,6;MQ=60;FQ=-60    GT:PL:GQ        1/1:189,33,0:63
I       2914    .       G       A       96      PASS    DP=12;VDB=0.0280;AF1=0.5;AC1=1;DP4=2,2,2,5;MQ=60;FQ=57;PV4=0.58,0.0066,1,0.43   GT:PL:GQ
I       3022    .       C       G       23      VDB     DP=5;VDB=0.0135;AF1=0.5;AC1=1;DP4=2,1,1,1;MQ=60;FQ=26;PV4=1,1,1,1      GT:PL:GQ        0/1
I       3106    .       T       A       15.1    VDB     DP=7;VDB=0.0135;AF1=0.5;AC1=1;DP4=3,2,1,1;MQ=60;FQ=18.1;PV4=1,0.29,1,1  GT:PL:GQ        0/1
I       3197    .       G       A       26      VDB     DP=7;VDB=0.0112;AF1=0.5;AC1=1;DP4=3,0,1,3;MQ=60;FQ=28.2;PV4=0.14,6.3e-06,1,1    GT:PL:GQ
I       3226    .       G       A       9.52    Qual;VDB        DP=8;VDB=0.0112;AF1=0.5;AC1=1;DP4=2,2,1,2;MQ=60;FQ=12.3;PV4=1,9.5e-08,1,1      GT:
I       3688    .       G       A       21      PASS    DP=9;VDB=0.0240;AF1=0.5;AC1=1;DP4=2,2,1,3;MQ=60;FQ=24;PV4=1,1.7e-08,1,1 GT:PL:GQ        0/1
```

- How many samples are included in the vcf?
- Did site 3106 pass the filters? If not, why? And site 2825?
- What type of variant is site 2875?
- What is the genotype at position 2709? And what is its genotype quality?

variant calling – samtools mpileup

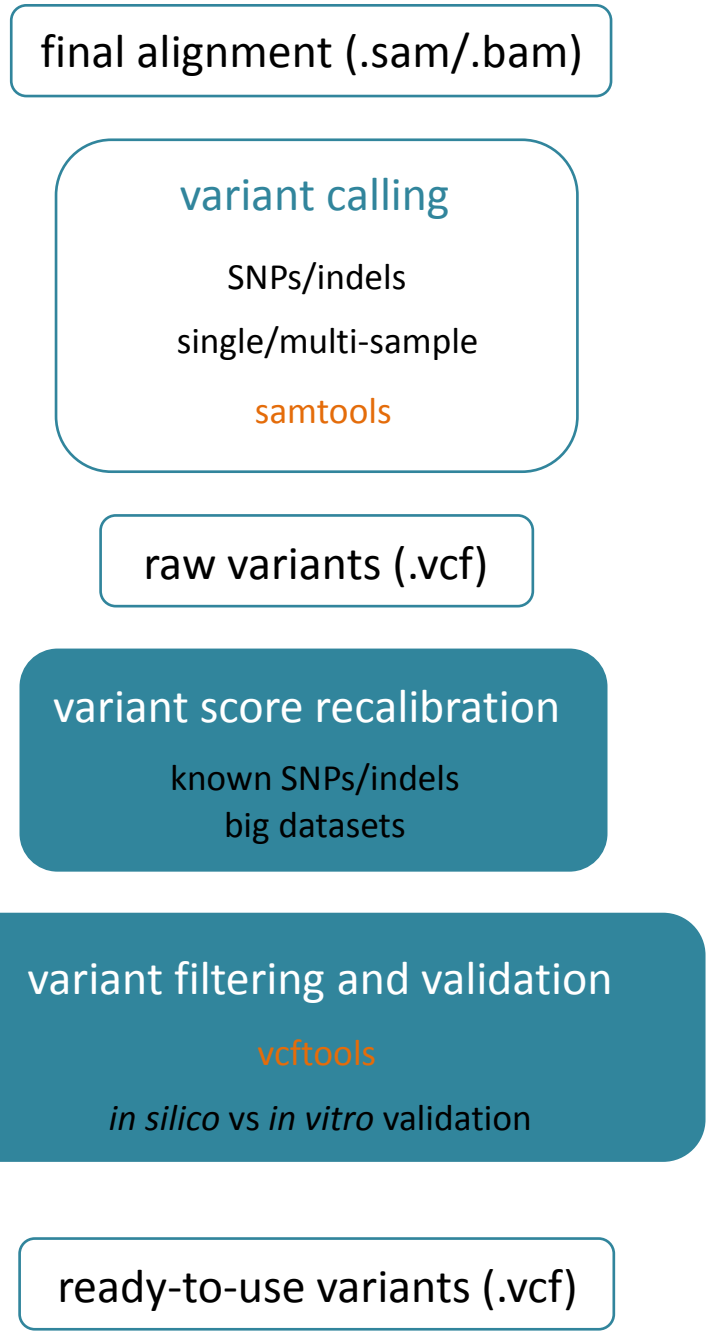
```
samtools mpileup -u -Q 20 -q 50 -g -s -f
Saccharomyces_cerevisiae.EF4.68.dna.toplevel.fa
library_final_sorted.bam | bcftools call -mv  - >
variants_raw.vcf
```

What is the meaning of the options for samtools mpileup? And for bcftools view?

Check your vcf

```
more variants_raw.vcf
```

What if you use the option -t in samtools mpileup? What's the difference you observe in the two vcf files?


final alignment (.sam/.bam)
variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
variant filtering and validation
vcftools
in silico vs *in vitro* validation
ready-to-use variants (.vcf)

variant calling – filtering variants

Common cautions:

- Base quality	BQ20
- Depth (min and max)	very dependent on your average
- Mapping quality	MQ50/60
- Strand-bias	p-value>0.05
- SNP density	dependent on the genome [e.g. no more than 1 SNP/4bp]
- Indel proximity	not closer than 10bp to an indel

Keep in mind your project may have some specific requirements

For example, if you are studying homologous regions (or you are using a distant reference genome), which is the parameter you should tailor first?

Some filters may be applied during the variant calling while others are applied afterwards

Further reading: “Consensus Rules in Variant Detection from Next-Generation Sequencing Data” Jia et al 2012 PLoS One

variant calling – variant quality score recalibration

Available in GATK.

It aims at producing well-calibrated probabilities for the variants called.

It develops a continuous, co-varying estimate of the relationship between SNP call annotations (e.g. MQ, QD…) and the probability that a SNP is a true genetic variant versus a sequencing or data processing artefact.
It needs “true sites” to be trained.

We are not going to use it, because it needs big datasets (either many samples, or whole genome data) to work properly.
You can find more information at
http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html

variant calling – vcftools

http://vcftools.sourceforge.net/docs.html

It not only allows to filter variants but it includes all sorts of useful options to handle your vcf files and extract useful information out of it

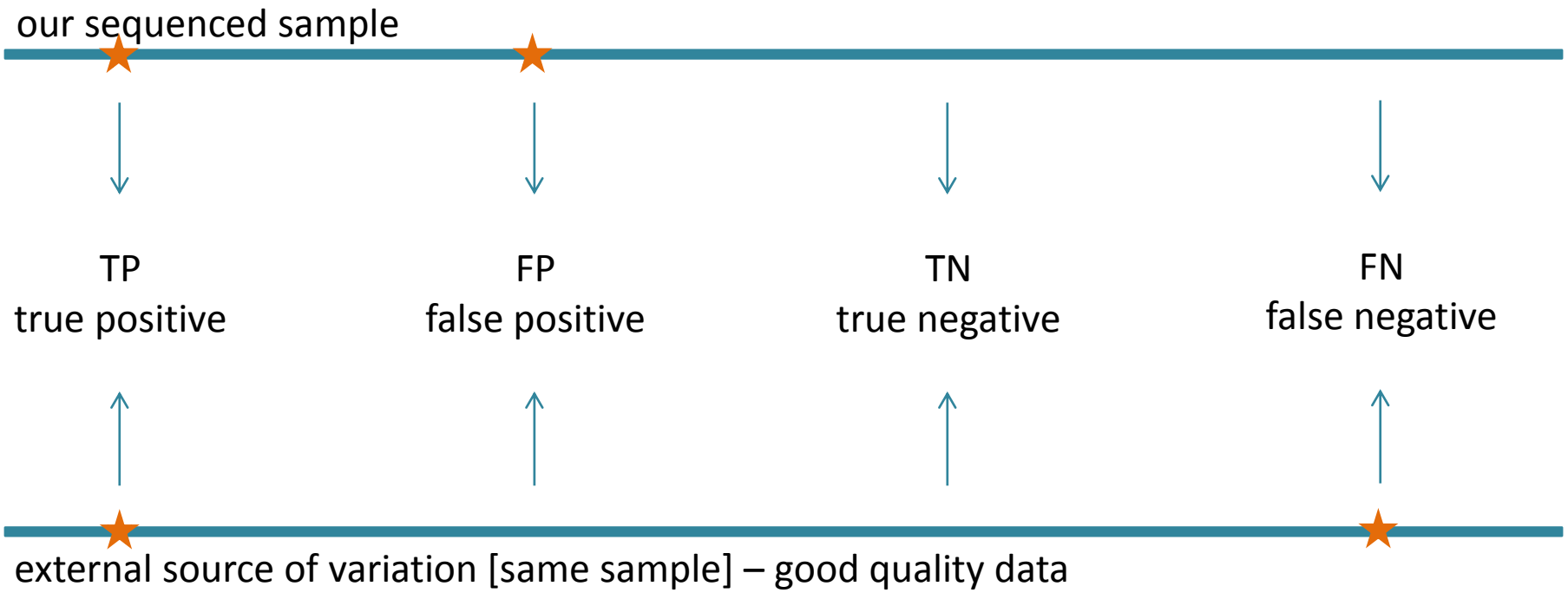
```
module load vcftools
cat variants_raw.vcf | vcf-annotate -f d=2/w=10 >
variants_flt.vcf

more variants_flt.vcf
```

If you consider the first ten variants, how many did not pass the filters applied? And why?

variant calling – evaluating

Specificity vs Sensitivity = False Positive vs False Negative



high specificity → low FP

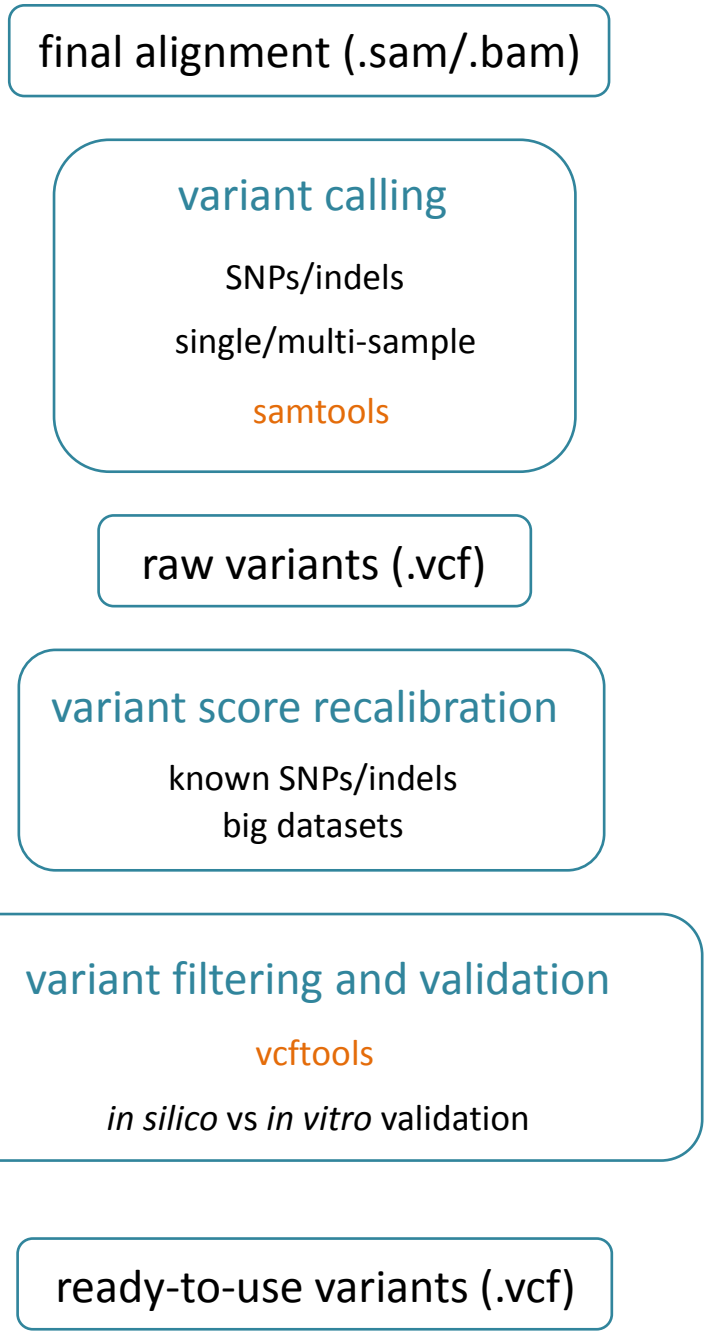
high sensitivity → low FN

variant calling – validation

external source of known variation – sequencing a sample for which you have independent data will help to understand the quality of your data (also reducing the need for experimental validation)

experimental validation – select a number of newly discover variants to be tested with a different technology (usual sanger sequencing); the rate of false discovery will give an estimate of how well the sequencing performed

lack of standards for validation rates and acceptable false discovery rates


final alignment (.sam/.bam)
variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
variant filtering and validation
vcftools
in silico vs *in vitro* validation
ready-to-use variants (.vcf)